Data aggregations for facilities science

"Linking datasets and articles for publication",
Harwell, 30 April 2013

Vasily Bunakov and Brian Matthews
(STFC Scientific Computing Department)

STFC Scientific Computing Department

Facility Data Archives

All **ISIS** data (~25 years) > 3,000,000 files

All **Diamond** Data (~5 years) > 100,000,000 files

CERN LHC Tier 1 Data

UK hub for **LHC** data (~3 years: 11PB)

Computing Architectures:

1 - the UK's most powerful computer (IBM BlueGene/Q :1.4 petaflops)

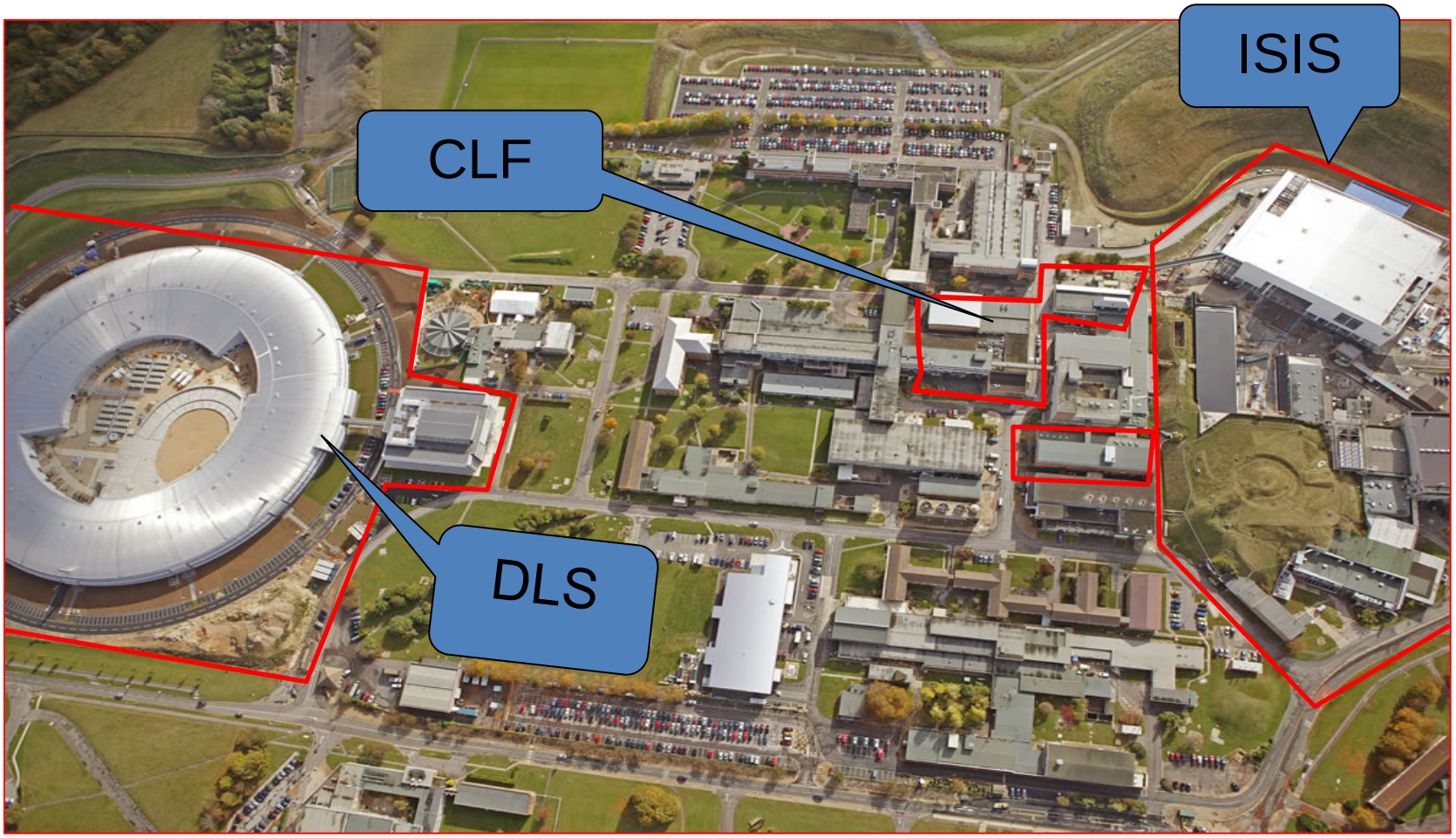
2 - the UK's most powerful graphics processor computer (190,000 graphics cores : 248 teraflops)

3- large commodity computing server (7000 processor cores)

4 - high throughput super-data-cluster (4.6 petabytes of parallel file storage with 1 terabit per second aggregate bandwidth from the data to the processors)

The StorageTek tape robot 100PB Capacity

Facilities Support



Big Facilities for Small Science

Facilities Data Lifecycle

As in PanData-ODI – D6.1 (which has much more detail)

Proposal

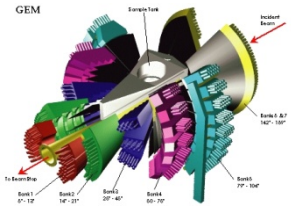
Scientist submits application for beamtime

Approval

Facility committee approves application

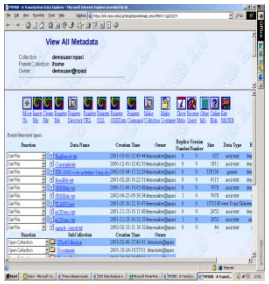
Scheduling

Facility registers, trains, and schedules scientist's visit



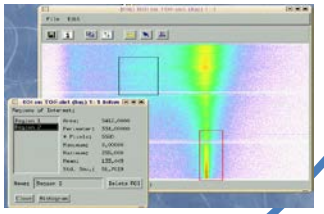
Experiment

Scientists visits, facility run's experiment



Data storage

Raw data filtered, and stored



Data analysis

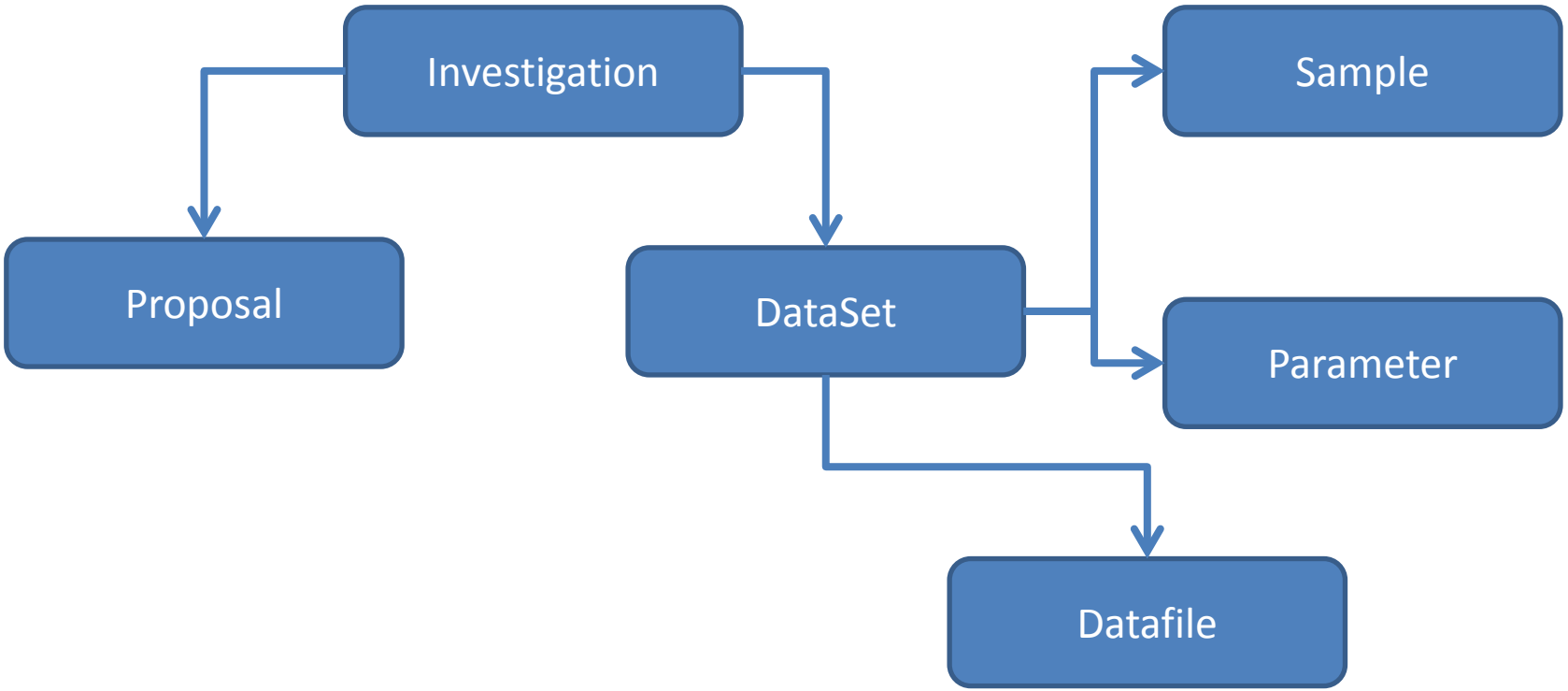
Tools for processing made available

Record Publication

Subsequent publication registered with facility

http://code.google.com/p/icatproject/

Simplified Metadata Model

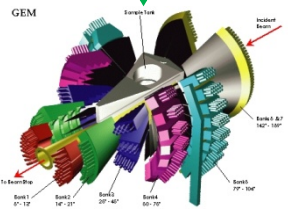


After proposal: Initialise the Research Object

:investigator

Investigation #n
DOI:STFC.xxx.n

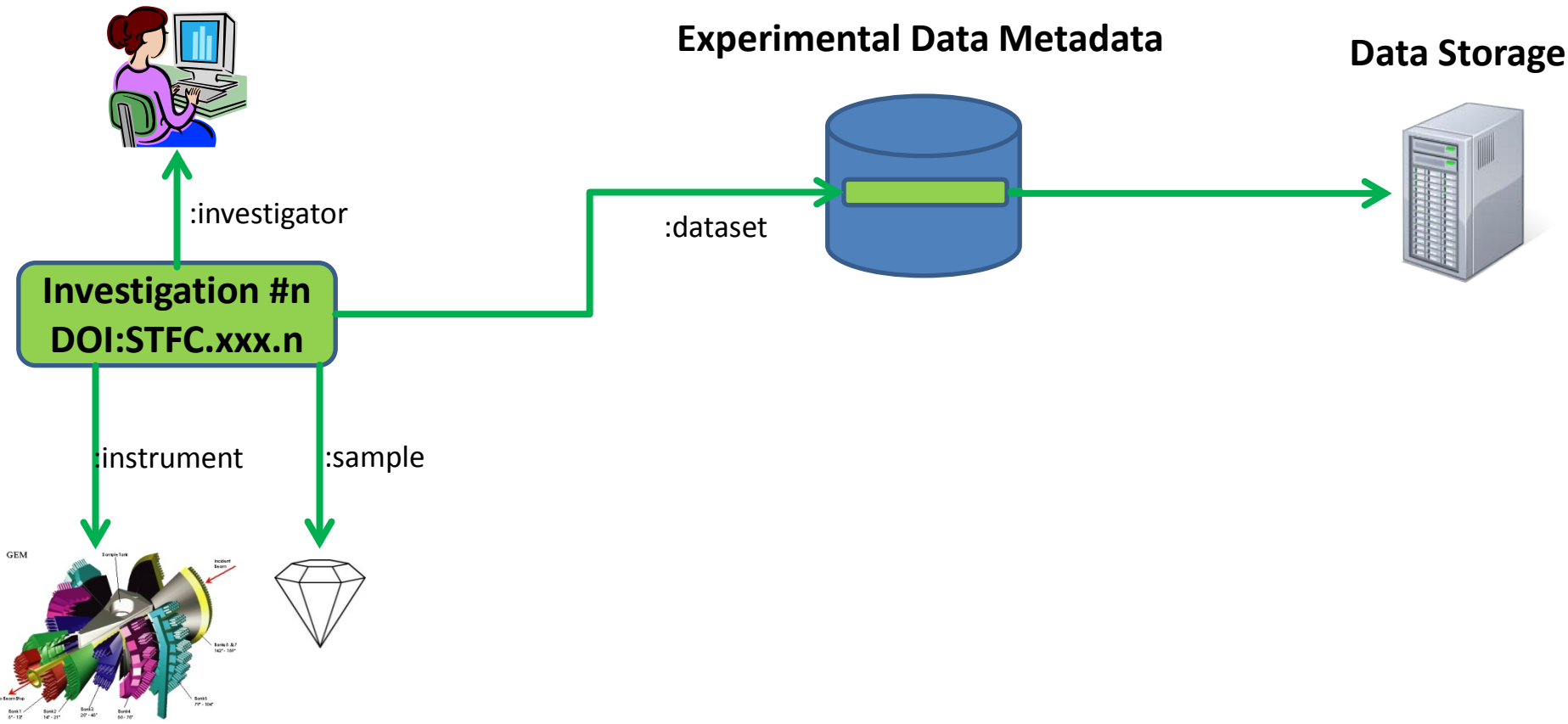
:instrument




```
:n a  csmd:Investigation ;
        csmd:investigation_doi  doi:stfc.xxx.n
        csmd:investigation_investigationUser   :iu1 ;
        csmd:investigation_instrument  :inst1 .

:iu1  a csmd:investigationUser ;
        csmd:investigationUser_user  :u1 .

:u1 a csmd:User .

:inst1 a csmd:Instrument  .
```

- Assign (but not necessarily register) a DOI for the object
- Take basic investigator and instrument information from the proposal system
- Also link to funding

After the experiment



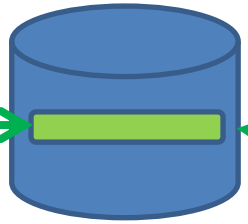
- Own metadata format (CSMD)
- More or less what ICAT currently supports
- Adds extra details on parameters, datasets, formats etc.

Linking the Publications

:investigator

Investigation #n
DOI:STFC.xxx.n

:dataset



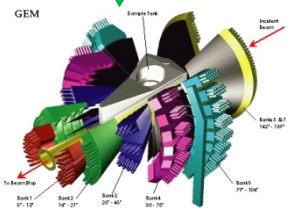
Experimental Data Repository

cito:cites

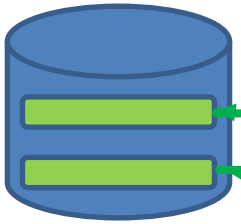
cito:cites

:instrument

:sample



Publication Repository



Publication Store

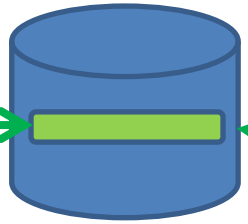
- Own metadata format (CSMD)
- cito – citation ontology (Oxford)
- Would also need to take into account pub metadata
- Publication repository could be on a different site

Linking Publication into Research Object

:investigator

Investigation #n
DOI:STFC.xxx.n

:dataset



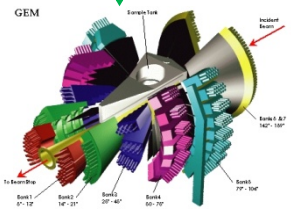
Raw Data Repository

cito:cites

cito:cites

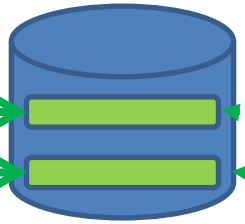
:instrument

:sample



:publication

:publication



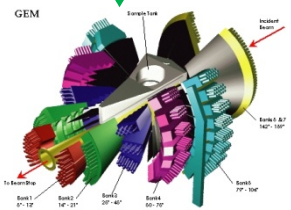
Publication Repository

Linking the derived data into the Research Object

:investigator

Investigation #n
DOI:STFC.xxx.n

:instrument



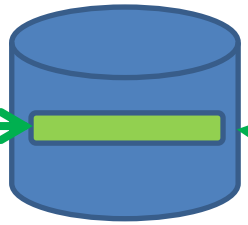
:sample

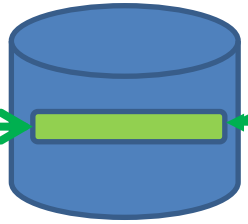
:dataset

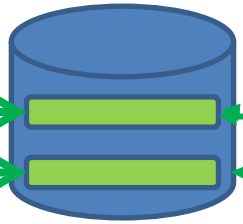
:relatedDataset

:publication

:publication







cito:cites

cito:cites

cito:cites

cito:cites

cito:cites

- Own metadata format (CSMD)
- OAI-ORE
- Cito

Our DataCite entries are in fact Investigations

(red is for “data” notion, and green is for “investigation”)

Data collected on the GEM instrument at the ISIS facility

ISIS Data

RB920025.

Investigation title: Crystal and magnetic structures of EuWO1+xN2-x.

Creator: *Kusmartseva,A*
Creator: *Rodgers,J A*
Creator: *Attfield,J P*

DOI: 10.5286/ISIS.E.24071239

Date of Experiment: Tue Aug 04 14:38:23 BST 2009

Publisher: STFC ISIS Facility

Data format: RAW/Nexus
Select the data format above to find out more about it.

See the next slide

DOWNLOAD
download the dataset

Data Citation

The recommended format for citing this dataset in a research publication is as:
[author], [date], [title], [publisher], [doi]

For Example:
Kusmartseva,A. et al; (2009): 920025, STFC ISIS Facility, doi:10.5286/ISIS.E.24071239

Abstract

Eu2+ d0- transition metal perovskites are of interest as potential multiferroics when undoped, or as CMR materials. EuWO1+xN2-x is a new magnetoresistive material and exists over a broad range of x = -0.2 to 0.5. It has a ferromagnetic ordering transition at TC = 12 K. Neutron diffraction is needed to determine the I112/m monoclinic superstructure evidenced by TEM that arises from O/N ordering and octahedral tilting, and the magnetic order. This may include a coexistence of antiferromagnetic/ ferromagnetic orders (as found in a previous GEM study of the analogueEuNbO2N). 2 days on GEM are needed to study 2 samples with different x values (one stoichiometric x =0, the other highly doped x = 0.5) because of high absorption by Eu.

“DataCite abuse”

As we have seen, we use DataCite for Investigations, with Datasets only referred from them

Other data curators sometimes use DataCite for Publications (“documents”) that contain data: http://data.datacite.org/10.7480/OA

So “data” DOIs tend to resolve either into Investigations or Publications

Publication and Investigation similarity

Feature / aspect	Publication	Investigation
Is an intellectual entity	V	V
Is a subject of peer review	V	V (via proposal approval)
Can cite other intellectual entities	V	V
Can be cited by other intellectual entities	V	V
Citation chains (steps of discourse) observed	V	V
Universal identifiers “mints” available	V	V

This gives Investigation a potential for a “full membership” in the research discourse along with Publication.

Datasets and software are likely to remain “associated members” because of absent feature 3 and de-facto weaker features 2 and 6.

Basic principles of building research objects for facilities science

- Follow research lifecycle
- Consider Investigation a RO “seed”
- Apply Linked Data principles
- Re-use existing vocabularies and ontologies
- Share ROs via recognizable data formats and APIs

Problems and challenges

- Universal IDs for Investigations are still a novelty: there is not many of them
- Lack of IDs for other components: instruments, experimental techniques, …
- Proto-objects most circulate within a “native” facility (although PANDATA raise hopes)
- Many researchers, data practitioners, publishers and policy makers are unaware of the potential of Research Objects as intellectual entities

Some day in future

"Facility-relevant" clouds in a larger Research LOD cloud powered by ontologies